



# A Credit Card Default Prediction Method Based on CatBoost

Yikai Zhao<sup>(✉)</sup>

National University of Singapore, Queenstown, Singapore  
ikaiz@u.nus.edu

**Abstract.** This paper presents a study on the prediction of credit card user default using the CatBoost model. The dataset used in this study is a credit card dataset from a financial institution. The dataset contains information about the credit card users such as their age, gender, credit limit, and payment history. The CatBoost model was used to predict the probability of default for each user. The results showed that the CatBoost model was able to accurately predict the probability of default for credit card users. And in the experiment, I found that the prediction effect of CatBoost model is better than that of XGBoost, Lasso, and LightGBM. The results of this study can be used to help financial institutions better manage their credit card portfolios and reduce the risk of default.

**Keywords:** credit default · CatBoost · feature engineering · machine learning

## 1 Introduction

People's consumption intentions and economic consciousness have been encouraged by the improvement of the material economy level. Because of its large profit margin and widespread acceptance, the credit card has become one of the primary ways for financial institutions to generate profits. For regular users, it is quick and simple, and features like delayed payment make their life easier. Yet, the issue of credit card infractions is growing more and more prevalent due to the high rise in credit card users and the recent deterioration of the economic situation. Banks and other financial institutions will suffer significant losses if the necessary actions are not implemented in time for control. As a result, it's important to assess the credit standing of credit card holders and apply reliable, scientific techniques to forecast default situations.

This paper presents a machine learning approach using the CatBoost to help credit card companies predict customer default. The proposed method uses desensitized tagged data to analyze customer data and identify patterns that can be used to predict customer default. The results of the proposed method are compared with traditional methods to demonstrate its effectiveness.

Sections 2 describes the associated work. In Sects. 3 and 4, we provide an overview of our procedures and tests.

## 2 Related Work

Since the introduction of neural networks in 1990 [1], Support Vector Machine in 2010 [2], and more recently ensemble methods like Random Forests (RF), the variety of techniques used for business credit scoring and bankruptcy prediction has greatly increased [3]. However, the emergence of these new tools gradually brought about a brand-new problem in the shape of a contentious efficiency explainability tradeoff. These new algorithms' statistical results do, in fact, significantly outperform those of the more established methods, but because to their inner complexity—often referred to as a “black box,” it is impossible to confidently explain the judgments they generate. Although recent studies have addressed the issue by either developing explanation models [4] or combining the best of both worlds in efficient and understandable new techniques [5], this problem became even more difficult as clients and financial regulators stressed the need for clarity and explainability of the scoring processes.

Machine learning is an important subfield of artificial intelligence that comes from statistical model fitting. Machine learning combines inference and sample learning to create the appropriate theory from the facts, especially when working with “noisy” patterns and large data sets. It is becoming more important for analyzing large samples, many vectors, and confusing data.[6]. The main characteristics of machine learning methods are as follows: machine learning can get the results that best fit the data by continuously learning and looping to optimize the target problem; the fitted model can help us explore more statistical relationships between data system characteristics and variables and can find more complex patterns in the data [7]; and machine learning uses a variety of methods, such as regularization and pruning, to better solve the overfitted. The prediction of stock returns [8]. Sami Ben Jabeur et al. used LightBGM to predict the oil price during the COVID-19 epidemic [9]; Liu, Yingr et al. used LightBGM to study whether Digital Inclusive Finance can predict household wealth and analyze the characteristics of strong predictive ability for household wealth.Sami Ben Jabeur et al. Used CatBoost to predict bank bankruptcy[10]. Fujimoto Shouji et al. used CatBoost to interpolate the non-random missing values in the big data of financial statements [5].

## 3 Methodology

The Russian search engine company Yandex created and released the CatBoost algorithm, a machine learning algorithm based on the Gradient Boosting Framework, in 2017. The benefit of CatBoost is that it can be used to issues including classification, regression, and sorting. It also performs exceptionally well in terms of model training time, accuracy, and resilience. CatBoost can adaptively process categorical features, which are typically one of the bottlenecks of conventional GradientBoosting algorithms, and its primary purpose is to address large-scale classification issues.

The CatBoost method differs from other tree-based gradient lifting algorithms in the following ways: category features are added during the creation of the decision tree structure, and the tree structure is created by combining with other features, i.e., by crossing the category features with other numerical features in order to better recognize and utilize the data in the category features; CatBoost employs many models. In order to

increase forecast accuracy and aid in avoiding over-fitting, it will incorporate different models. After each training session, CatBoost will automatically add a new model. This novel model can enhance the performance of predictions for data points with significant mistakes.

In summary, the CatBoost algorithm is a versatile, quick, and accurate algorithm that is simple to use.

## 4 Experiment

The experiment data used in this study comes from the Kaggle competition for the American Express - Default Prediction. The goal of the competition is to forecast, based on a customer’s monthly customer profile, the likelihood that they won’t pay off their credit card bill in the future.

The target binary variable is derived by tracking performance over the 18 months following the most recent credit card statement, and a default event is deemed to have occurred if the consumer does not make the required payment within 120 days of the statement date. Each customer’s aggregated profile characteristics at each statement date are contained in the dataset. Features fall into the following broad groups after being anonymised and normalized. Variable prefixes and types are shown in Table 1.

### 4.1 Feature Engineering

In order to enhance the feature set, we have carried out some feature engineering. First of all, for multiple pieces of data of the same user, I will aggregate them together as one piece of data to participate in the calculation. Then, we generate some statistical characteristics for default rate, expenditure, payment, balance and risk characteristics to increase the entire feature set. In addition, function combination creates additional functions, such as payment and date functions, by combining multiple functions. Reduced memory and training acceleration require feature selection. I show the main process of the entire feature engineering in Fig. 1. The feature importance histogram is used to select important features in Fig. 2.

**Table 1.** Variable prefix and type

Variable Prefix	Variable type
D_*	Delinquency variables
P_*	Payment variables
S_*	Spend variables
B_*	Balance variables
R_*	Risk variables

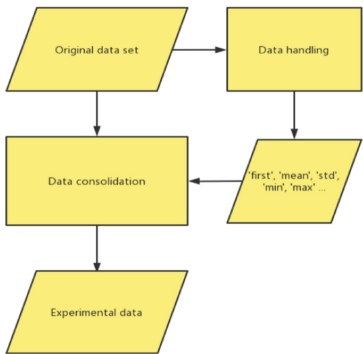


Fig. 1. Flow chart of feature engineering

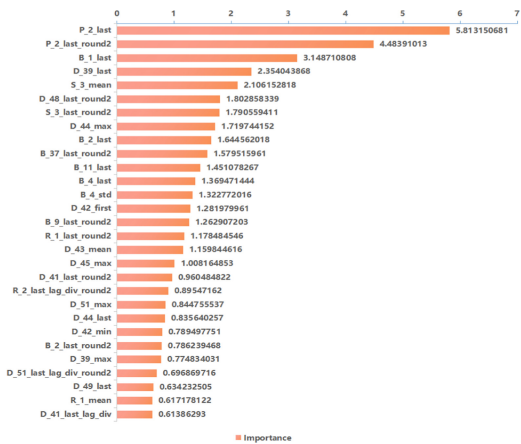


Fig. 2. Importance of features analyzed by CatBoost model

4.2 Training Parameters

According to the empirical method and grid search method, the parameters of CatBoost are obtained. We selected the training parameters shown in Table 2.

4.3 Evaluation Metrics

The evaluation index M is the average of two ranking measures: the normalized Gini coefficient G and the default rate D is 4%.

$$M = 0.5 \cdot (G + D)$$

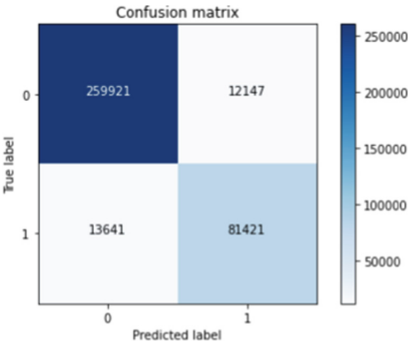
The default rate captured at 4% is the percentage of the positive labels (defaults) captured within the highest-ranked 4% of the predictions, and represents a Sensitivity/Recall statistic. For both of the sub-metrics G and D, the negative labels are given a weight of 20 to adjust for down sampling.

**Table 2.** Training parameters

Parameter	Value
n_estimators	3000
learning_rate	0.0845
depth	6
border_count	254
l2_leaf_reg	3
bayesian_matrix_reg	0.1

**4.4 Evaluation Results**

We specially demonstrated the Confusion matrix of classification results of the CatBoost model in Fig. 3. In order to evaluate our experimental performance, we did compare several mainstream algorithms. The higher the metric, the better the model. The experimental results are shown in Table 3. The CatBoost algorithm we used has the highest 0.799 metric in these models, 0.005, 0.011 and 0.03 higher than Xgboost, LightGBM and Lasso respectively.



**Fig. 3.** Confusion matrix of classification results

**Table 3.** Experimental results for each model

Models	Metric
Xgboost	0.794
Catboost	0.799
LightGBM	0.789
Lasso	0.769

## 5 Conclusion

We did feature engineering and used CatBoost as our model to predict credit card default. We introduce the relevant work and algorithm model in Sect. 2 and Sect. 3 respectively. In the Sect. 4, we present our experiment. We introduce some methods of feature engineering and give the specific parameters of the model. In the experimental part, our model CatBoost has the highest 0.799 among these models, 0.005, 0.011 and 0.03 higher than Xgboost, LightGBM and Lasso respectively.

**Acknowledgement.** I would like to express my heartfelt gratitude to Prof. Jiao, my mentor, who provided a lot of help and guidance in writing this article. His professional knowledge in machine learning and enthusiasm for teaching inspired me to write this article, and I thank him for his help and encouragement.

## References

1. Adner, R., Puranam, P., & Zhu, F. (2019). What Is Different About Digital Strategy? From Quantitative to Qualitative Change. *Strategy Science*, 4(4), 253–261. <https://doi.org/https://doi.org/10.1287/stsc.2019.0099>
2. Athey, S., & Imbens, G. W. (2017). The State of Applied Econometrics: Causality and Policy Evaluation. *The Journal of Economic Perspectives*, 31(2), 3–32.
3. Ben Jabeur, S., Khalfaoui, R., & Ben Arfi, W. (2021). The Effect of Green Energy, Global Environmental Indexes, and Stock Markets in Predicting Oil Price Crashes: Evidence from Explainable Machine Learning. *Journal of Environmental Management*, 298, 113511. <https://doi.org/https://doi.org/10.1016/j.jenvman.2021.113511>
4. Crook, J. N., Edelman, D. E., & Thomas, L. C. (2005). Credit Scoring. *The Journal of the Operational Research Society*, 56(9), 1003–1005. <https://doi.org/https://doi.org/10.1057/palgrave.jors.2602037>
5. Fujimoto, S., Mizuno, T., & Ishikawa, A. (2022). Interpolation of Non-Random Missing Values in Financial Statements' Big Data Using CatBoost. *Journal of Computational Social Science*, 5(2), 1281–1301. <https://doi.org/https://doi.org/10.1007/s42001-022-00165-9>
6. Jabeur, S. B., Gharib, C., Mefteh-Wali, S., & Ben Arfi, W. (2021). CatBoost Model and Artificial Intelligence Techniques for Corporate Failure Prediction. *Technological Forecasting and Social Change*, 166, 120658. <https://doi.org/https://doi.org/10.1016/j.techfore.2021.120658>
7. Kim, H. S., & Sohn, S. Y. (2010). Support Vector Machines for Default Prediction of SMEs Based on Technology Credit. *European Journal of Operational Research*, 201(3), 838–846. <https://doi.org/https://doi.org/10.1016/j.ejor.2009.03.036>
8. Malekipirbazari, M., & Aksakalli, V. (2015). Risk Assessment in Social Lending via Random Forests. *Expert Systems with Applications*, 42(10), 4621–4631. <https://doi.org/https://doi.org/10.1016/j.eswa.2015.02.001>
9. Odom, M. D., & Sharda, R. (1990). A Neural Network Model for Bankruptcy Prediction. In 1990 IJCNN International Joint Conference on Neural Networks (pp. 163–168 vol.2). <https://doi.org/10.1109/IJCNN.1990.137710>
10. Sezer, O. B., & Ozbayoglu, A. M. (2018). Algorithmic Financial Trading with Deep Convolutional Neural Networks: Time Series to Image Conversion Approach. *Applied Soft Computing*, 70, 525–538. <https://doi.org/https://doi.org/10.1016/j.asoc.2018.04.024>

11. Fujimoto, S., Mizuno, T., & Ishikawa, A. (2022). Interpolation of non-random missing values in financial statements' big data using CatBoost. *Journal of Computational Social Science*, 5(2), 1281-1301. doi: <https://doi.org/10.1007/s42001-022-00165-9>

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

